Charging and Billing Issues in High Speed Heterogeneous Networking Environments

Marion Raffali-Schreinemachers, KPN Research, Royal PTT Nederland N.V.
Sathya Rao, Telscom, Switzerland
Frank Kelly, Lyndewode Research, Great Britain (Annex)
Yannis Markopoulos, Intrasoft, Greece.

Abstract: Charging is a requirement to recover the costs of network resources employed to deliver services. In broadband networks charging at the ATM level is necessary for customers of a native ATM service, and as a basis for charging higher layer services using ATM for transport. For variable bit rate traffic charging is expected to reflect the usage of network resources by the traffic. However, resource allocation for variable bit rate traffic is difficult given only the ATM traffic contract. The theory of effective bandwidth more closely describes the resource needs of a traffic profile. However, it relies on extensive knowledge of the traffic characteristics, far beyond the parameters of the traffic contract. An approximation to the effective bandwidth can be used for resource allocation. This approximation of the effective bandwidth is based solely upon the mean rate of the connection, and its traffic contract. A possibility to obtain an estimate of that mean rate is to ask the user or his application to provide it. An incentive to the user to provide a correct estimate, is to use the estimate also in charging. The charge is calculated through the effective bandwidth approximation based on the estimated mean rate, and by the realised mean rate measured for the connection. This mechanism benefits both network operator and customer. The network can allocate its resources more sparingly while still fulfilling its service obligations. The customer pays less for service as less resources can be allocated. He pays the least possible charge if his real mean corresponds closely to his estimate. Also, the measurements required in the implementation of the charging mechanism are simple enough to be included in networks.

1. Introduction

Broadband networks offering broadband services need to include facilities for charging and billing. In this context, charging designates the calculation of a charge for a connection. The charge is calculated based on some characteristics of the connection, according to a charging scheme, which in turn is part of a tariffing policy. Billing encompasses the process that registers the characteristics of calls, and combines them with charging information to yield a bill to communicate to the customer. In both processes, different parties can act as the customer. Such possible customers are end-users, service providers, or network operators who make use of services offered by other operators or providers.

In the heterogeneous broadband networks operating and planned at present, ATM plays an important role as a transport layer technique. ATM-level services are offered to customers who have native ATM end-users, and to service providers who rely on ATM to transport other broadband services. The networks are heterogeneous at the ATM-level as they carry a diverse mixture of constant and variable bit rate connections, with very

Proceedings Interworking '96; Nara (Japan), October 1-3, 1996; pp. 97-108.

different rates and quality requirements. Allocating resources and charging such diverse connections poses a great challenge to network operators and service providers. Concepts of effective bandwidth offer a means to deal with the diverse connections and their requirements. Effective bandwidth approximations can be shown to be simple to implement, while very effective in capturing closely resource allocation requirements.

When broadband services based on ATM are offered and charged, charging at the ATM level is necessary. Native ATM customers are charged directly at the ATM level, while customers of higher layer services based on ATM are charged at those higher service layers, by charging that takes into account the charge incurred by the service at the ATM level.

This paper presents results and issues of charging at the ATM level investigated and obtained in the ACTS project CA$hMAN. It discusses different approaches to charging in conjunction with resource allocation and usage in networks. Ample attention is given to effective bandwidth theory as a means to quantify resource usage. A review of the theory in the annex to the paper serves as background to the charging issues. The requirements that different parties pose on charging and billing, and the receptiveness of users and customers to the incentives intended by the charging schemes are discussed. First results of implementation and user response to an effective bandwidth charging schemes are available from CA$hMAN.

2. Charging at the ATM Level

Native ATM services serve customers who provide services that rely upon ATM for transport, or to end-users operating native ATM applications and equipment. In such cases the provider of the ATM-level service will charge his customer at the ATM-level. In the case where an ATM service is not delivered between parties, but rather used within an operator to transport his own higher layer service, ATM-level charging principles may serve to describe the cost of transporting the service over ATM in the network. Especially in a network combining several operators and techniques, such indications of the impact of the service on the network resources may help to charge the end-to-end service over the conglomerate network.

Charging is closely related to other processes in the network. Resource allocation and connection admission control use much of the same information relevant for charging. Such information encompasses the traffic contract for the connection, and traffic profiles. It stands to reason, that the burden that a traffic stream places upon the network and its resources, determine both whether it will be admitted to the network, as well as how much it will be charged for its use of the resources. Similarly, resources reserved for the exclusive use by a given customer, may well be charged to that customer, not because he used them, but because he rendered them unusable for others. Thus if the allocation of resources can be performed sparingly, i.e. with sufficient margin to guarantee the service agreed in the traffic contract, but with as little allocation as possible beyond the real usage, then effective resource usage can be maximised, more connections can be admitted, and for equal revenue, each or some connections can be charged less.

The question how sparingly resources can be allocated to a connection depends greatly on the information available describing the traffic that is to be carried over that connection. If only the parameters of the ATM traffic contract are available, resource allocation must allow for worst case behaviour on the part of the traffic stream, while still ensuring the guaranteed QoS (Quality of Service). For a DBR (Deterministic Bit Rate) connection that amounts to Peak Cell Rate (PCR) allocation, taking into account also the CDVT (Cell Delay Variation Tolerance). For an SBR (Statistical Bit Rate) connection, worst case requirements can be deduced from the declared PCR, SCR (Sustainable Cell Rate), CDVT and IBT (Intrinsic Burst Tolerance) or the MBS (Maximum Burst Size). For variable bit rate traffic streams these worst case allocations are factors greater than the allocation required had the traffic profile been known exactly in advance. Concepts developed to approximate more closely the resource requirements of a traffic stream rely on more information than only the parameters of the traffic contract, without assuming complete prior knowledge of the traffic profile. Effective bandwidth is used in some such concepts for resource allocation as well as for charging.

An important assumption in the following is that usage based charging closely match the resource allocation policy of the network. Thus, usage describes the resources required to guarantee the service on a connection, rather than a pure cell count. While a user may initially consider only correctly received cells as part of the service rendered, he is also sensitive to the argument that some margins have to be provided and charged to cater for incomplete predictability.

3. Requirements on Charging and Billing

Different parties in internetworking pose requirements to charging. Some of these requirements, relevant to usage based charging and customer receptiveness, are exposed here. Other requirements such as legal conditions and regulatory issues need to be considered in the strategy of operators setting charging policy.

Under the assumption that resources are scarce, the usage of network resources by a service should be reflected in the charge. In this manner, charging provides an incentive to use resources sparingly. To provide this incentive, even the basic information provided in the ATM traffic contract for a connection may suffice. For instance, in a network that allocates resources and charges connections according to the PCR of the contract, customers will understand that they can lower their charges by decreasing their PCR, without necessarily affecting the total volume of traffic carried over the connection, or even the perceived quality of the service. An example is a multimedia-conferencing application that operates in a very bursty fashion with a PCR of 12 Mb/s. Shaping its output to a smoother stream with 8Mb/s PCR may have no visible or audible effect on the quality, and noticeable effect on the charge.

Whatever charging scheme is operated, registration of traffic contracts, measured parameters and charge must be cost-effectively implementable. Charging and billing can be huge cost factors in networks, so that the cost of a service may come to depend largely on the cost of charging and billing, rather than on the resource usage at the ATM level. It is important to guard at an early stage against approaches to charging and billing that will entail top-heavy implementations. Implementing a resource allocation scheme and associated charging that ensure a small but significant optimisation in effective resource usage, while entailing a large and significant increase in accounting and billing costs, makes little business sense.

Finally in the realm of the operator, charging needs to be compatible across network operator boundaries. Two such boundaries can be identified as follows. The charging of higher layer services needs to cover the charge of the ATM transport below. If charging at the ATM level and above are not compatible, incentives are given to service providers to look elsewhere for their transport. If the provider of the service and of the ATM network are not the same party, the service provider will look for another network if the ATM charging is prohibitive to his higher layer service. And if the higher layer service is too expensive in relation to the ATM transport, competing service providers will undercut the service operator's prices. A second boundary for charges is between different networks at the transport level, ATM or otherwise. Here parameters for usage and charging of ATM level resources need to be agreed and exchanged, and translated in case of interworking between different transport techniques. Charging schemes need not be shared, standardised or public, but the parameters that they rely on must be standardised and exchanged, to allow unambiguous charging.

More requirements to charging stem from customers, be they end-users, service providers or network operators in a multi-operator environment. To customers, charging needs to be clear, traceable, understandable, and to some degree predictable. Not necessarily the charging structure or the tariffing policy need to be transparent, but the choices to be made by the customer, and the implications of the alternatives to him must be understood.

If the customer or user is required to provide information to the system about his intended traffic profile, he needs to be presented with an adequate interface. If he has a choice of tariffs, these must be made clear, and the differences must be traceable and visible, either through a network application, or in communication between the customer and the provider's sales force.

Customers who request an advice of charge either before, during or after the connection want to rely upon a meaningful estimate, understandable and fitting to their requirements. Billing also needs to be reliable and understandable, containing all and only relevant information. Thus the network has to register and retain this information. This process requires mechanisms and resources.

Many requirements need consideration in the design of a charging scheme. Some of them determine properties of the scheme, while other ones prohibit certain implementations. The models and the charging scheme described in the next section do not solve all challenges posed by the above requirements. However care has been taken to infringe upon none.

4. Models for Charging

Usage based charging intends to link the charge for a connection to the network resources necessary to service the connection according to its traffic contract. Charging for usage makes sense if it bases its assessment of resource usage upon the same or similar criteria as the connection admission process that allocates the resources. If resource allocation is based on the PCR of the traffic contract, then the amount of resource made unavailable for other connections is described by that PCR, so that a compatible charge would also take the PCR as a basis for resource usage. In both the areas of resource allocation and charging both network operator and customer aspire to a scheme more closely reflecting the resources effectively used and necessary for the connection, and thus to a usage charge closer to the usage perceived by the customer.

To attain a better estimate of resources required and charge incurred, more information about the traffic profile needs be included in the resource assessment, in addition to the traffic parameters from the ATM traffic contract. In the extreme, if all were known about the traffic profile, its required resources could be very closely determined, with only marginal over-allocation. However, complete knowledge of the traffic profile is available only after its passage, thus no longer serving a meaningful purpose in predicting requirements. The network needs a mechanism that will predict resource requirements, and that can be operated with limited additional information. Also the network must be able to obtain helpful estimates of that limited information at call establishment, when resources need to be allocated. Effective bandwidth theory provides means to describe the resource requirements of a connection with some precision, based upon some elements of the traffic profile. In the instance described here, the mean rate of the connection will serve as additional information to the traffic contract parameters.

Figure 1 gives an overview of the approximations involved. Section 7 of this paper gives more detailed background to the concept of effective bandwidth. Figure 1 illustrates concepts of actual usage and approximations to it for a connection with a given traffic contract and thus a given PCR. The resource usage of the connection is mapped against its mean rate in the graph. The mean rate is determined as the total volume of cells transferred during the connection, averaged over the connection time. In the figure, the lower dotted area indicates the actual usage of resources; this quantity cannot be realistically determined in a live networking environment, and differs for connections that do have a same contract and mean rate. The dashed line above the real usage is the effective bandwidth of the connection based on complete knowledge of the traffic profile; this quantity is also difficult to obtain in a realistic environment. The solid curve above represents an approximation of the effective bandwidth as a function of the traffic contract parameters and the mean rate of the connection; it has a number of desirable properties. It is an upper bound to the effective bandwidth based on the complete traffic profile, and it is strictly concave in the mean rate M. This will be essential in the following, where an estimate of the mean is used as a prediction of the real mean for a connection.

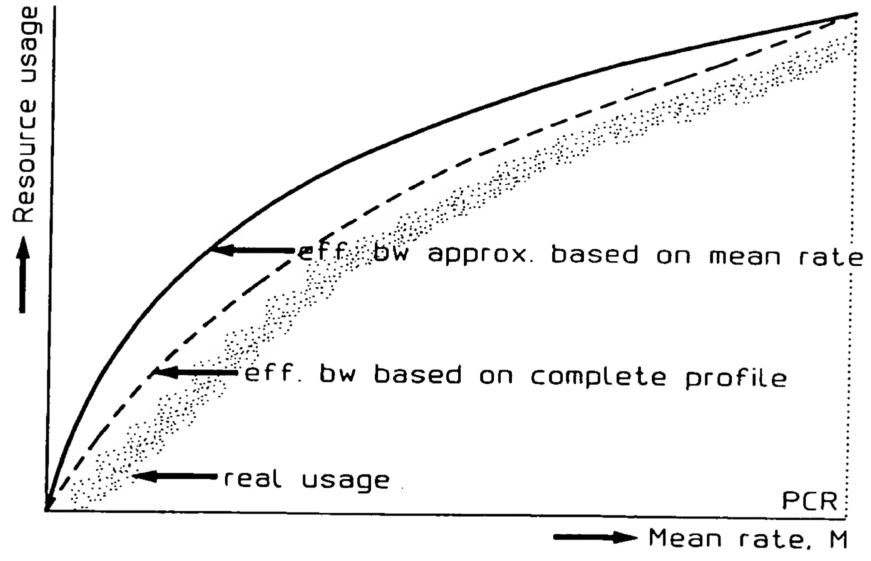

Figure 1: Resource usage and effective bandwidth.

If the mean of a connection were known along with its traffic contract, at the establishment of a call, the effective bandwidth approximation based on the mean could guide network resource allocation with significant savings in comparison to PCR allocation. As for the other elements of the traffic profile, an exact value of the mean is available only at the end of the connection for any variable bit rate traffic stream. Thus the network needs a reliable estimate of the mean rate at call establishment. That estimate can be based on registrations performed by the network or some assisting application for similar traffic over time, or could be obtained from the customer (or his application or equipment) who could employ a similar historic registration of mean rates for traffic types. If the customer provides the estimate for his mean rate, charging can provide some incentive to the customer to provide as accurate an estimate as possible. Figure 2 depicts the mechanism of this incentive.

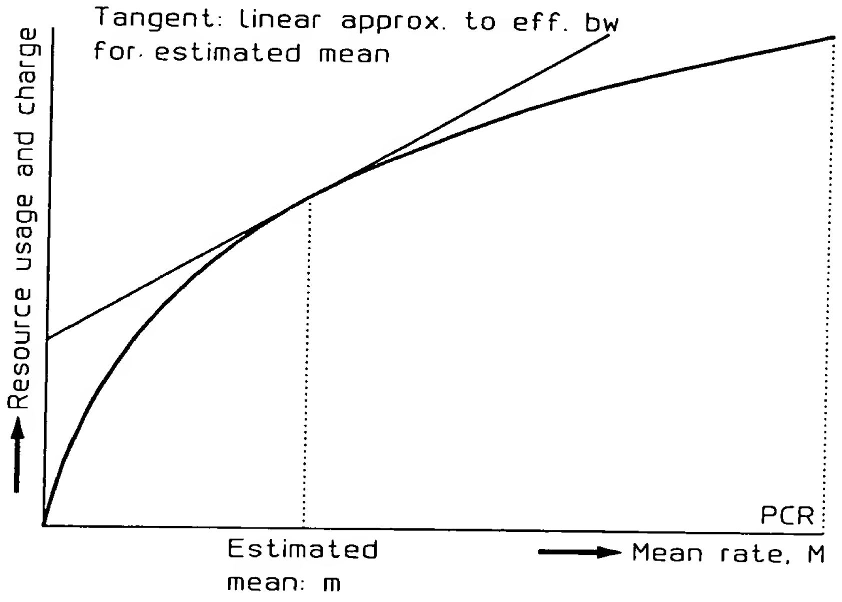

Figure 2: Charge as an upper bound to network resource usage.

The formula is given as a(m)*T + b(m)*V + c(m), where T is the connection time, and V is the volume carried over the connection. The parameters a, b and c in the expression giving the charge are determined by the estimated mean that fixes the tangent's intersection with the effective bandwidth curve. They also convey to what extent a connection is charged for volume, and to what extent for time. a(m) May be interpreted as the charge for a unit of time, b(m) as the charge for a unit of volume, and c(m) as a connection set-up charge. The set-up charge c(m) has little relation to resource usage. At the estimated mean m, a tangent to the curve that approximates effective bandwidth based on the mean, describes the charging formula for the connection. The a(m), b(m) and c(m) place the tangent on the curve at the estimated mean m. The expression can be formulated in terms of the realised mean rate M of the connection and the connection time T, as V/T = M. As the curve is strictly concave, the charging function yields a charge larger than the network resource usage for any realised mean that is different from the estimated mean. Thus the customer pays the least possible charge for his connection if he accurately predicts its mean. In that case he has also provided most assistance to the network.

In the charging formula, the choice of a triple a,b,c depends on the estimated mean m. The possible values of these triples need not be continuous: finite granularity will suffice, as the reliability of the estimates is bounded. A choice of a finite number of tangents, three for instance, may be enough. The format in which such a choice is presented to customers or other sources of estimated means must suit the customer's perception and understanding. A possible expression of the choice could ask: Is the expected mean less than half the PCR, more than three-quarters of the PCR, or somewhere in between? A last possibility: would you like to be charged mostly on volume, mostly on time, or in a balanced way for both? User trials and interviews will give more insight into preferable expressions to present the choice to different customers or sources.

Two remarks remain in the discussion of approximations to effective bandwidth. First, if resource allocation is based on an estimated traffic profile (such as characterised by the estimated mean) the network cannot rely on the fact that the estimation gives an upper bound for the required resources. Extra margins may have to be included in the allocation. Second, the approximation to the effective bandwidth based on the real mean of the connection can be enhanced to be closer an approximation to effective bandwidth through the inclusion of more information, for instance to capture more precisely the burstiness of connections. Whenever information elements are added, the improvement in resource allocation, and the extra complexity and required accuracy have to be carefully weighed.

5. Management and Signalling Implications of Charging

Collecting information for charging and registration of traffic needs to be performed in the network. Information from the traffic contract, and possibly estimates and measured parameters need to be transferred from the UNI (User Network Interface), the customer, and the measurement devices respectively, to the charging unit in the network. Means need be devised for that transport. If charging in the sense of multiplying measured parameters and charging parameters (a,b,c (m)) to obtain a charge is to be performed at the management centre in a network, management channels may be considered for the transfer of these parameters from UNI and measurement devices to the management centre. For the estimates obtained from the user, signalling might be altered to include this additional information with the signalled traffic contract parameters.

Also, if an advice of charge is to be presented to a customer, an end-system or a network element, the advice of charge needs to be transferred around the network. Again signalling or management channels may be possible implementations. Whatever the means chosen to implement charging, they will claim network resources in their operation. Also the load on the network to accommodate the registration and transfer of billing information in the network may be considerable.

Finally, in a multi-operator environment, information related to charging, such as measured parameters and choice of charging parameters, needs to be agreed and exchanged, and possibly translated across network boundaries. Standardisation of such parameters, their meaning and their means of transferral through the network is a prerequisite to usage based charging in multi-operator environments. ITU-T SG13 will initiate work in area in its next study period.

6. Merits of the Charging Mechanism and Customer Perception

Reviewing the requirements to charging and the merits of the charging mechanism described here, encouraging conclusions can be reported. First, using the same approximation to effective bandwidth for both resource allocation and charging, ensures that the usage of resources by the service is reflected in the charge. Also this allocation offers a mayor improvement over allocation relying only on traffic contract parameters.

Second, initial implementations of the charging mechanism have shown it to be implementable in a cost-effective manner. The parameters time and volume of a connection that need to be measured during a call are already included in most network equipment available today. The other parameters, a(m), b(m) and c(m) are constants chosen from a file or table. No processing need be dedicated to them as part of network operation. Also, the additional parameters to be registered for a call for billing and administration are few: time, volume and tariff choice (a,b,c-triple).

Also, for network boundaries between operators, the charging parameters to be agreed and exchanged are clear: time and volume in addition to the traffic contract.

Finally, the charging formula can be explained in a simple manner, in terms of charge per unit volume, and cost per unit time. It also guarantees that the charge incurred by a connection never exceeds the charge for that connection with resources allocated according to peak cell rate. Thus, both network operator and customer benefit. The network saves resources by employing the approximation to effective bandwidth. Thus it may be able to support more connections in parallel. The customer is charged less for a same service if he estimates his mean and thus enables the network to save resources. The least charge to the customer is achieved if his estimate corresponds to the realised mean.

Beside the unavoidable objective of recovering costs for network resources employed to deliver the broadband service, charging may serve additional objectives. Two aims discussed above are to provide incentives: to encourage users to submit traffic more easily catered for by the network, on the one hand, (such as smooth traffic with a PCR as low as possible, for instance), and on the other hand to encourage users to announce traffic profiles in reliable detail so as to assist the network in its resource allocation.

User trials are performed in ACTS project CA$hMAN to investigate the response and receptiveness of users and customers to such incentives. In this context the distinction between customers and users is essential as they have different priorities, and may therefore respond differently to an incentive. The customer who provides networking facilities to his employees is interested in the charge incurred. The user, who relies upon the network to accomplish his tasks, may be less sensitive to alterations in bills that he neither sees nor foots. For a user, the trade-off between quality and charge, for instance through traffic shaping, is determined by his perception of quality, rather than his possibly vague perception of cost. Whatever the incentives, the intended effects, and the user groups concerned, their effect largely depends on the quality of the user interfaces, and the clarity of the communication between the provider and the customer organisation.

Annex: Effective Bandwidths, a Background to Charging Models.

This section reviews the concept of an effective bandwidth, beginning with a simple model. Suppose that J sources share a single unbuffered resource of capacity C, and let X_j be the load produced by source j. Assume that $X_j, j = 1,\dots,J$, are independent random variables, with possibly different distributions. Can the resource cope with the superposition of the J sources? More precisely, can we impose a condition on the distributions of $X_1,\dots,X_J$ which ensures that

$$P\left\{\sum_{j=1}^{J} X_j > C\right\} \le e^{-\gamma} \tag{1}$$

for a given value of γ? The answer to this question is, by now, fairly well understood.

There are constants s, C' (depending on g and C) such that if

$$\sum_{j=1}^{J} \alpha(X_j) \le C' \tag{2}$$

where

$$\alpha(X_j) = s^{-1} \log \mathrm{E}e^{sX_j} \tag{3}$$

then condition (1) is satisfied. The expression (3) is called the *effective bandwidth* of source j. This result, originally due to Hui [1], was generalized in [2] to show that if the resource has a buffer, and if the load produced by source j in successive time periods is a sequence of independent bursts each distributed as X_j then the probability the delay at the resource exceeds b time periods will be held below $e^{-\gamma}$ provided inequality (2) is satisfied, with α again given by equation (3), where $C' = C$ and $s = \gamma / (bC)$.

It is by now known that for quite general models of sources and resources it is possible to associate an effective bandwidth with each source such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee (see [3, 4, 5, 6, 7]). Often the relevant definition is of the form

$$\alpha(X_j) = (st)^{-1} \log \mathrm{E}[e^{sX_j[0,t]}] \tag{4}$$

for particular choices of s and t, where $X_j[0,t]$ is the load produced by source j over an interval of length t. There may be several constraints of the form (2) corresponding to different physical or logical resources within a network.

Consider the very simple case of an on/off source of peak rate h and mean rate m, for which the value for the parameter t in (4) is chosen small compared to the length of times for which the source is typically on and off. The exact choice of t is determined by the amount of buffer at the network ([3]), and a reasonable choice could be the time it takes for a burst of the source to fill the buffer. Let

$$P\{X = 0\} = 1 - \frac{m}{h}, P\{X = h\} = \frac{m}{h} \tag{5}$$

The effective bandwidth (4) of such a source is then

$$\alpha(h,m) = \frac{1}{st}\log\left[1+\frac{m}{h}(e^{sth}-1)\right] \tag{6}$$

For fixed h this function is increasing and concave in m, while for fixed m it is increasing and convex in h. As s tends to 0 (corresponding, in this example, to a very large capacity C in relation to the peak h), the effective bandwidth approaches m, the mean rate of the source. However as s increases (corresponding to a larger peak h in relation to the capacity C) the effective bandwidth increases to the peak rate h of the source.

Charging Mechanisms and User Information - The on-off Fluids

One possible charging mechanism might measure the effective bandwidth of a connection, perhaps by estimating expression (3) using an empirical averaging to replace the expectation operator. There is, however, a simpler *indirect* mechanism [8, 9, 10], which has important additional advantages in the co-ordination of information and characterization effort between users and the network.

To illustrate the mechanism, consider first the case of an on/off source with a known (and possibly policed) peak rate h, but with a mean rate that may not be known with certainty, even to the user responsible for the source. Assume, however, that the user has a prior distribution G for the mean rate M of the call. The distribution G may represent very vague information, or might be constructed by recording past observed mean rates. Then the expected mean rate of the call is

$$\mathrm{E}_G M = \int_0^h x dG(x) \tag{7}$$

If the network knew the prior distribution G for the mean rate M, then the network would determine the effective bandwidth of the call, from equations (3) and (7), as

$$\frac{1}{st}\log \mathrm{E}\, e^{sX[0,t]} = \frac{1}{st}\log \mathrm{E}_G \mathrm{E}(e^{sX[0,t]} | M) = \frac{1}{st}\log \mathrm{E}_G\left[1+\frac{M}{h}(e^{sth}-1)\right]$$

$$= \frac{1}{st}\log\left[1+\frac{\mathrm{E}_G M}{h}(e^{sth}-1)\right] \tag{8}$$

But expression (8) is just the effective bandwidth if M is not random, but identical to its mean value under G. We see that since the source is on/off with known peak rate the network need only know $E_G M$, the user's expected mean rate; further detail about the distribution G does not influence the effective bandwidth, and would be superfluous for the network to even request. How, then, should the network encourage the user to assess and to declare the user's expected mean rate? We next investigate whether the charging mechanism might be used to provide the appropriate amount of encouragement.

Suppose that, before a call's admission, the network requires the user to announce a value m, and then charges for the call an amount $f(m; M)$ per unit time, where M is the measured mean rate for the call. We suppose that the user is risk-neutral and attempts to select m so as to minimize $E_G f(m; M)$ the expected cost per unit time: call a minimizing choice of m, m' say, an *optimal* declaration for the user. What properties would the network like the optimal declaration m' to have? Well, first of all the network would like to be able to deduce from m' the user's expected mean rate $E_G M$, and hence the effective bandwidth (8) of the call. A second desirable property would be that the expected cost per unit time under the optimal declaration m' be proportional to the effective bandwidth of the call (or, equivalently, *equal* to the effective bandwidth under a choice of units). In [8, 9] it is shown that these two requirements essentially characterize the tariff $f(m; M)$ as

$$f(m; M) = a(m) + b(m)M \tag{9}$$

defined as the tangent to the curve $\alpha(M)$ at the point $m = M$ (see Figure 3 where $B(h, M)$ stands for $\alpha(M)$ with the parameter h symbolizing the dependence on the peak rate).

The key property used in the proof [9] is the strict concavity of $\alpha(M)$ as a function of M. By a simple differentiation of the function (7), the coefficients in expression (9) are given by

$$b(m) = \frac{e^{sh} - 1}{s[h + m(e^{sh} - 1)]}, \quad a(m) = \alpha(m) - mb(m) \tag{10}$$

where the dependence of the coefficients on the peak rate h is now made explicit.

Note that a very simple interpretation is available for the tariff *f*: the user is free to declare a value m, and then incurs a charge $a(m)$ per unit time and a charge $b(m)$ per unit of volume carried.

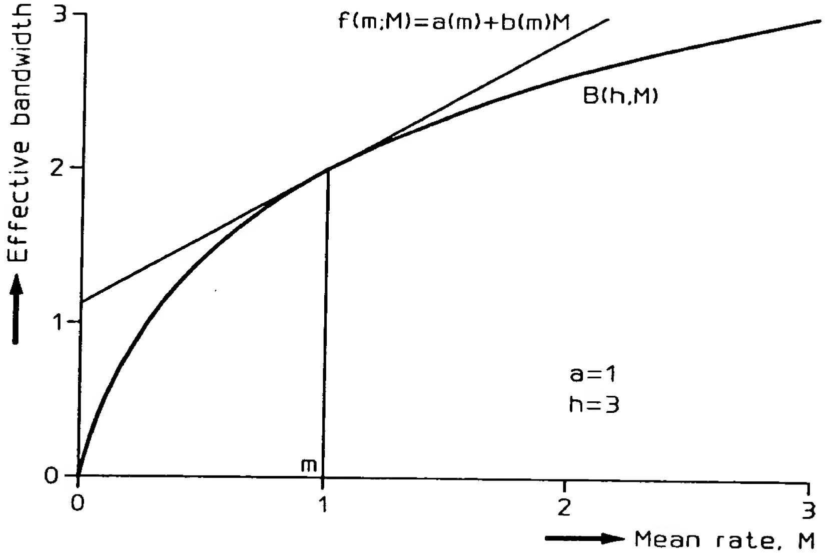

Figure 3: The tariff.

The properties characterizing the tariff f have many interesting and desirable consequences. For example, suppose that a user can, with some effort, improve its prediction of the statistical properties of a call. A crude method of deciding upon the declaration m might be to take the average of the measured means for the last n calls, but more sophisticated methods are possible. If the user is an organization containing many individuals, the user might observe the identity of the individual making the call, the applications active on that individual's desktop computer, as well as the called party, and utilize elaborate regression aids to make the prediction m. Is it worth the effort? In [8, 9] it is shown that improved prediction reduces the expected cost per unit time of the connection by exactly the expected reduction in the effective bandwidth required from the network. This is an important property: users should not be expected to do more work determining the statistical properties of their calls than is justified by the benefit to the network of better characterization.

References

[1] J.Y. Hui. Resource allocation for broadband networks. *IEEE J. Selected Areas in Commun.* 6, 1598—1608, 1988.

[2] F.P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems* 9, 5—16, 1991.

[3] C. Courcoubetis, and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. To appear in *Journal of Applied Probability,* 1996.

[4] G. de Veciana, and J. Walrand. Effective bandwidths: call admission, traffic policing and filtering for ATM networks. *Queueing Systems, 1994.*

[5] G. Kesidis, J. Walrand, and C. S. Chang. Effective Bandwidths for Multi-class Markov Fluids and other ATM Sources. *IEEE/ACM Trans. Networking*, 1(4):424-428, August 1993.

[6] A. Elwalid, and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking*, June 1993.

[7] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems* 2, 71—167,1993.

[8] F.P. Kelly. On tariffs, policing and admission control of multi-service networks. *Operations Research Letters* 15 , 1—9, 1994.

[9] F. P. Kelly. Tariffs and effective bandwidths in multi-service networks. In J. Labetoulle and J.W. Roberts, editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress - ITC 94*, volume 1a of Teletraffic Science and Engineering, pages 401-10. Elsevier Science B. V., June 1994. Antibes Juan-les-Pins.

[10] F.P. Kelly. Notes on Effective Bandwidths. *In Stochastic Networks: Theory and Applications,* Oxford University Press, 1996.